Folding of pig gastric mucin non-glycosylated domains: a discrete molecular dynamics study

Bogdan Barz · Bradley S. Turner · Rama Bansil · Brigita Urbanc

Received: 8 June 2012 / Accepted: 27 July 2012 / Published online: 28 September 2012
© Springer Science+Business Media B.V. 2012

Abstract Mucin glycoproteins consist of tandem-repeating glycosylated regions flanked by non-repetitive protein domains with little glycosylation. These non-repetitive domains are involved in polymerization of mucin and play an important role in the pH-dependent gelation of gastric mucin, which is essential for protecting the stomach from autodigestion. We examine folding of the non-repetitive sequence of PGM-2X (242 amino acids) and the von Willebrand factor vWF-C1 domain (67 amino acids) at neutral and low pH using discrete molecular dynamics (DMD) in an implicit solvent combined with a four-bead peptide model. Using *the same implicit solvent parameters*, folding of both domains is simulated at neutral and low pH. In contrast to vWF-C1, PGM-2X folding is strongly affected by pH as indicated by changes in the contact order, radius of gyration, free-energy landscape, and the secondary structure. Whereas the free-energy landscape of vWF-C1 shows a single minimum at both neutral and low pH, the free-energy landscape of PGM-2X is characterized by multiple minima that are more numerous and shallower at low pH. Detailed structural analysis shows that PGM-2X partially unfolds at low pH. This partial unfolding is facilitated by the C-terminal region GLU236-PRO242, which loses contact with the rest of the domain due to effective “mean-field” repulsion among highly positively charged N- and C-terminal regions. Consequently, at low pH, hydrophobic amino acids are more exposed to the solvent. In vWF-C1, low pH induces some structural

Electronic supplementary material The online version of this article (doi:10.1007/s10867-012-9280-x) contains supplementary material, which is available to authorized users.

B. Barz · B. Urbanc (✉)
Physics Dept., Drexel University, Philadelphia, PA 19104, USA
e-mail: brigita@drexel.edu

B. S. Turner
Division of Gastroenterology, Beth Israel Deaconess Medical Center and Harvard Medical School,
Boston, MA 02215, USA

R. Bansil
Physics Dept. and Center for Polymer Studies, Boston University, Boston, MA 02215, USA

changes, including an increased exposure of CYS at position 67, but these changes are small compared to those found in PGM-2X. For PGM-2X, the DMD-derived average β -strand propensity increases from 0.26 ± 0.01 at neutral pH to 0.38 ± 0.01 at low pH. For vWF-C1, the DMD-derived average β -strand propensity is 0.32 ± 0.02 at neutral pH and 0.35 ± 0.02 at low pH. The DMD-derived structural information provides insight into pH-induced changes in the folding of two distinct mucin domains and suggests plausible mechanisms of the aggregation/gelation of mucin.

Keywords Mucin • PGM-2X • vWF-C1 • Mucin gelation • DMD simulation • Protein folding • Free-energy landscape • Pig gastric mucin

1 Introduction

Mucin is the major glycoprotein component of the mucus layer that coats the exposed epithelial surfaces of organs and is responsible for the protective and lubricative functions of mucus [1, 2]. This complex molecule resembles a multiblock copolymer with the hydrophilic polysaccharide-coated protein brushes alternating with somewhat hydrophobic protein domains with little or no glycosylation. Recombinant DNA techniques have shown a striking similarity in the basic structure of the underlying mucin genes [3]. All of the mucin gene products contain domains comprising tandem-repeating sequences that are rich in the amino acids SER, THR, and PRO (*O*-glycosylation sites) located in the heavily *O*-glycosylated portion of the molecule. In all mucin apoprotein genes, this domain occupies the central region of the polypeptide backbone and is flanked by CYS-rich, non-repeating sequences at the N- and C-termini (see Fig. 1) [4]. These non-glycosylated regions resemble typical globular proteins in their amino acid composition and are involved in the polymerization of the primary subunit via disulfide linkages between the SH groups of CYS into a dimer. Further polymerization leads to polymers with molecular weight ranging from 2 to 20 million Da, of which 70–80% is sugar. These polymers further aggregate, and, in the case of the gastric mucin, gel under acidic pH [5–7], and protect the stomach from autodigestion [8].

In previous studies, Bansil, Turner et al. [1, 5–7, 9] examined the gelation of gastric mucin at acidic pH values using light scattering, fluorescence binding, atomic force microscopy, and rheology. For gastric mucin gels below $\text{pH} = 4$, these experiments showed that enzymatic or biochemical disruption of the non-glycosylated regions of PGM prevents gelation and that hydrophobic regions exposed at low pH are involved in gelation, suggesting that gelation of PGM involves *the association of the non-glycosylated regions of the glycoprotein* [1], via a complex interplay between hydrophobic association and electrostatic interactions of charged amino acids. However, the molecular details of the gelation mechanism are not well understood. Using several different schemes for hydropathicity analysis, Cao [10] proposed that the PGM-2 X domain, which is interspersed between the glycosylated ST repeats (see Fig. 1), was a possible candidate for forming the crosslinks in a gel by hydrophobic association as it contained a hydrophobic pocket of 12–20 amino acids flanked by an ASP and ARG. The ionic interaction between the ASP and ARG would be disrupted below the pK of ASP ($\text{pK} \sim 4$) and this could potentially expose the amino acids in the hydrophobic pocket. Another possibility is that the aggregation involves the association of domains in the non-glycosylated C-terminal region bridged by the hydrophilic glycosylated regions, similar to gelation in a triblock copolymer.

Fig. 1 Sketch of porcine gastric mucin PGM showing different domains in the apoprotein (*top*), a S-S linked dimer (*middle*), and a multimer (*bottom*) illustrating the multi-block structure. Adapted from Dekker et al. [35]

To elucidate structural changes in mucin induced by low-pH conditions, we applied here discrete molecular dynamics (DMD) combined with a four-bead peptide model in an implicit solvent to study folding of two non-glycosylated domains, PGM-2X and vWF-C1, a portion of the vWF-C domain from the C-terminal region of the mucin apoprotein that were recently expressed in yeast [4] (see Fig. 1). The objective of this study is to characterize ensembles of folded structures of the two domains at neutral and low pH, quantify the pH-induced structural changes and thus elucidate the contribution of these two domains to the increased aggregation propensity of mucin at low pH. The conformational ensembles of 242-residue PGM-2X and 67-residue vWF-C1 are explored at neutral and low pH. Our analysis indicates that, whereas the vWF-C1 domain becomes more compact at low-pH conditions, the PGM-2X domain undergoes considerable conformational change and partially unfolds at acidic pH. This unfolding of the PGM-2X domain at acidic pH may contribute to mucin gelation to a much larger extent than the small pH-induced changes in vWF-C1 folding.

2 Materials and methods

2.1 Discrete molecular dynamics (DMD)

Discrete molecular dynamics (DMD) was introduced by Rapaport as a fast molecular dynamics method in which continuous interparticle potentials are replaced by a square-well

or a combination of square-well potentials [11]. The resulting dynamics is driven by collision events between pairs of particles that are otherwise moving along straight lines with constant speeds. During each collision, velocities of particles involved in the collision are updated while preserving the total kinetic energy, momenta, and angular momenta. Collisions (events) occur only at distances where the interparticle potential is discontinuous. In studies of proteins, DMD is typically combined with a coarse-grained protein model and effective interactions that mimic a solvent. Consequently, DMD is considerably faster than all-atom molecular dynamics (MD) with an explicit solvent, where all protein atoms are simulated together with thousands of water molecules. While studies of protein folding in all-atom MD are limited to time regimes of <1 ms, the DMD approach can be applied to study biologically relevant processes that occur on 10^5 fold longer time scales and involve hundreds of peptides [12].

2.2 Four-bead protein model and backbone hydrogen bond interactions

Introduction of a four-bead protein [13] model with backbone hydrogen bonding [14] in combination with DMD enabled an *ab initio* DMD approach, in which an explicit solvent is replaced by interactions mimicking an hydrophobic/hydrophilic effect and effective electrostatic interactions [15, 16]. Here, *ab initio* refers to the modeling approach that, unlike Go-like models, does not require the knowledge of the native state but need only the knowledge of the peptide sequence. In the four-bead protein model, there are three backbone atoms/beads corresponding to the amide N, the α -carbon C_α and the carbonyl C' groups. The side chain C_β is represented by one bead except for GLY, which lacks the side chain bead. The amino group of PRO is covalently bonded to its side chain C_β bead. In addition to covalent bonds within each amino acid and peptide bonds between adjacent amino acids, the model accounts for the specific protein geometry by introducing constraints between two adjacent amino acids [14]. Within each amino acid, the constraint of $2.44 \text{ \AA} \pm 0.5 \text{ \AA}$ between each amino group N and the corresponding side chain C_β are used to ensure the correct peptide geometry [14]. We modified this value for PRO and replaced it by $2.35 \text{ \AA} \pm 0.25 \text{ \AA}$ to account for its specific geometry. All lengths and standard deviations of all bonds and constraints are phenomenologically determined by using known crystalline structures of proteins from the Protein Data Bank (<http://www.rcsb.org/pdb/home/home.do>).

Among all naturally occurring amino acids, PRO can be found more often in a *cis* conformation than other amino acids. We can estimate a probability of *trans* to *cis* transition in PRO by analyzing the $C_\alpha[X]-C_\alpha[\text{PRO}]$ distance (X is the amino acid directly preceding PRO) and the dihedral angle ω for PRO using $\sim 15,000$ protein structures reported in the Protein Data Bank. Our data show that less than 5% of PROs adopts the *cis* conformation (Fig. S1, Supplementary Material). Because the PGM-2X domain contains 23 PROs, only one PRO on average is expected to adopt the *cis* conformation, whereas in the vWF-C1 domain with 10 PROs, less than one PRO is affected on average. We thus neglect the *trans* to *cis* transition in our DMD modeling approach.

Backbone hydrogen bonds are essential in any protein model as they stabilize α -helices and β -strands, basic secondary structure elements of proteins. In our model, each amino acid can form up to two backbone hydrogen bonds. As the carbonyl oxygen and the amide hydrogen are not explicitly present in the four-bead model, an effective backbone hydrogen bond is modeled between the nitrogen atom N_i of the *i*-th amino acid and the carbon atom C_j of the *j*-th amino acid [14]. The exception is PRO, which can form only one backbone hydrogen bond with its carbon group C_j but none with its amino group N_i . Four auxiliary

bonds are introduced to achieve the correct local geometry of the amino acids involved in the hydrogen bond formation: the two neighboring backbone atoms of N_i need to be at specific distances with respect to C_j and, conversely, the two neighboring backbone atoms of C_j need to be at specific distances with respect to N_i . The lengths and standard deviations of these auxiliary bonds as well as the hydrogen bond parameters are determined using a phenomenological approach in a similar way as used for the lengths and standard deviations of covalent and peptide bonds.

2.3 Residue-specific interactions due to hydrophobicity/hydrophilicity and charge

In our DMD approach, effective interactions between the side-chain atoms that mimic the solvent effects are implemented. Residue-specific effective interactions due to hydrophobicity and hydrophilicity [16] are based on the Kyte–Doolittle hydropathy scale [17]. When two hydrophobic side-chain beads are within the interaction range of 7.5 Å, they interact through a one-step attractive potential. When two hydrophilic side chain beads are within the same interaction distance, they interact through a one-step repulsive potential. In our model, the hydrophobic amino acids are ALA, CYS, PHE, LEU, MET, ILE, and VAL. The hydrophilic amino acids are ASP, GLU, HIS, LYS, ASN, GLN, and ARG. The side chains of the remaining amino acids GLY, PRO, SER, THR, TRP, and TYR interact only through excluded volume repulsion. We set the potential energy of the hydrogen bond, E_{HB} , which in proteins assumes a range of values ~ 1 –5 kcal/mol, to unit energy ($E_{HB} = 1$). The unit of temperature is E_{HB}/k_B where k_B is Boltzmann's constant and $T = 0.13$ was recently derived as an estimate of physiological temperature in the application of the DMD approach to amyloid β -protein ($A\beta$) folding relevant to Alzheimer's disease [18]. The absolute value of the maximal potential energy (strength) of the hydropathic interactions is set to $E_{HP} = 0.3$, which previously resulted in $A\beta$ oligomer size distributions that were consistent with *in vitro* data [19]. Note that by construction, the DMD model is generic and should not depend on the protein under study. The implicit solvent parameter E_{HP} , describing hydrophobic/hydrophilic effects, is expected to be similar in all aqueous solvents. We perform DMD simulations at constant temperature and volume using the Berendsen thermostat algorithm [20]. Although the Berendsen thermostat does not reproduce the canonical ensemble and is often replaced by the Nosé–Hoover method [21], it is used because of its computational efficiency in simulations of large systems that aim to find thermodynamic averages (rather than fluctuations). Effective electrostatic interactions between two charged side-chain beads are introduced in the model as an attractive (between oppositely charged beads) or repulsive (between beads of the same charge) two-step square-well potential with a cut-off at 7.5 Å. In addition to the charged side-chain beads, the N-terminal amino group is positively charged and the C-terminal carboxyl group is negatively charged, consistent with *in vitro* aqueous solution conditions. The potential energy of the effective electrostatic interaction E_{CH} is varied between 0 and 1 to explore the effect of E_{CH} on the folding of the two domains. Note that this implicit solvent parameter, which measures the relative strength of electrostatic interactions in the solvent, depends on ionic concentration.

2.4 Disulfide bonds

In addition to the four-bead peptide model parameterization [14], implementation of amino acid-specific interactions [15, 16] and PRO-specific parameters described above, we here

Fig. 2 The average β -strand propensity as a function of the implicit solvent parameter E_{CH} at neutral and low pH for DMD-derived folded structures of PGM-2X and vWF-C1. The error bars correspond to SEM values

also model disulfide bonding. Disulfide bonds formed between pairs of CYS amino acids are common to many proteins and help stabilize their native conformations. Different mucin domains are rich in CYS amino acids. We implement a disulfide bond as a simple square-well attractive reaction between two C_{β} beads of two CYS residues. In the Kyte–Doolittle hydrophathy scale, CYS is a hydrophobic amino acid and its C_{β} beads interact via attractive effective hydrophobic interaction at distances <7.5 Å. When two CYS C_{β} beads are at distances [3.6 Å, 4.2 Å], they form a disulfide bond with the potential energy $E_{DB} = 10.0$. This interaction range for disulfide bonding was chosen based on data reported by Dani et al., Fig. 2b [22]. Despite the large strength of disulfide bonding ($E_{DB} = 10.0$), disulfide bonds occasionally break and re-form, allowing for disulfide bond shuffling.

2.5 Units and implicit solvent parameters

We use the potential energy associated with one hydrogen bond, E_{HB} , as a unit of energy in our model. An application of the DMD approach to the A β folding study and a comparison to available experimental data resulted in an estimate of the physiological temperature of $T = 0.13$ [18]. Selecting $T = 0.13$ as the physiological temperature (310 K) and taking into account $k_B T = 0.6$ kcal/mol, we obtain $E_{HB} = 4.6$ kcal/mol or $7.6 k_B T$, which is within the expected range for protein interactions in aqueous solutions.

Taking into account the spatial resolution of 0.1 Å in our model and the unit energy $E_{HB} = 4.6$ kcal/mol, each simulation step corresponds to ~ 0.3 ps, which amounts to ~ 6 μ s for 20×10^6 simulation steps used as a total simulation time for vWF-C1 folding and ~ 18 μ s for 60×10^6 simulation steps used as a total simulation time for PGM-2X folding. It would be misleading to map these implicit solvent DMD-derived time scales directly to time scales measured *in vitro*. Due to the lack of collisions within the solvent and between the

solvent and the peptide, folding and assembly in the DMD approach occur much faster than *in vitro*. While the lack of an explicit solvent limits the DMD approach with respect to kinetic studies, it renders the approach computationally efficient enough to enable studies of folding and assembly of rather large proteins and peptides.

The two implicit solvent parameters, E_{HP} and E_{CH} , depend on the solvent of interest. For most proteins in an aqueous solution, E_{HP} is expected to be of a similar value. We thus use here $E_{HP} = 0.3$, which corresponds to $E_{HP} = 1.4$ kcal/mol or 2.3 k_BT, the value that was used in the A β studies of folding and oligomerization [16, 19, 23]. The solvent parameter associated with effective electrostatic interactions, E_{CH} , was varied between 0 and 1 (see Section 3.1) to explore the effect of ionic strengths on folding. In our structural analysis, only $E_{CH} = 0.8$ (or 3.68 kcal/mol) was used because it is in the expected range based on previously reported free-energy profiles between charged amino acids [24]). The value of 3.68 kcal/mol is approximately halfway between the values corresponding to the strongest electrostatic interactions of 4.5 kcal/mol reported for ARG-GLU and 2.5 kcal/mol reported for HIS-HIS.

2.6 DMD code with a four-bead model implementation

We used an in-house code for DMD simulations, which allows us to select the temperature, thermal coupling constant, and to modify other model parameters, such as the relative strengths of effective hydrophobic and electrostatic interactions, E_{HP} and E_{CH} and the system size in terms of both the volume and number of peptides. The code is flexible and can be modified to include new features. Importantly, we can select the option of studying the effect of each interaction individually by setting the other interaction strengths to zero. The four-bead peptide model with backbone hydrogen bonding was introduced by Ding et al. and all model parameters that are based on structural analysis of known protein crystalline structures from the Protein Data Bank were reported [14]. The exception is the N-C β bonding constraint for PRO, which was modified for the present study as described above. The residue-specific interaction scheme based on the Kyte–Doolittle hydrophathy scale and effective electrostatic interactions were implemented by Urbanc et al. as reported previously [15, 16]. Disulfide bond formation is newly implemented for the present study as described above.

2.7 Number of trajectories and total simulation time

The 242-residue PGM-2X and 67-residue vWF-C1 substantially differ in sequence length; thus, the number of trajectories and total simulation time are chosen for each of the domains separately. We acquired 32 trajectories of PGM-2X folding and 16 trajectories of vWF-C1 folding to achieve an efficient sampling of the phase space as described below. The total simulation time of 60×10^6 simulation steps (for PGM-2X) and 20×10^6 simulation steps (for vWF-C1) was chosen by monitoring the time dependence of the total potential energy for each trajectory such that the system is equilibrated within the final 10×10^6 simulation steps. Thus, all time frames between 50×10^6 and 60×10^6 simulation steps (for PGM-2X) and between 10×10^6 and 20×10^6 simulation steps (for vWF-C1), separated by 100,000 simulation steps, result in 3,232 folded conformations of PGM-2X and 1,616 folded conformations of vWF-C1 that were used for the structural analysis described below. All four-bead conformations were converted into the reduced-representation PDB

files that included, in addition to the four groups (N, C $_{\alpha}$, C and C $_{\beta}$), also the backbone oxygen (belonging to the C group) and hydrogen (belonging to the N group) needed for the backbone hydrogen bonding and analysis within the Visual Molecular Dynamics (VMD) software package [25].

2.8 Potential of the mean force (PMF)

We characterize each folded structure by a radius of gyration, R_g and a contact order, CO, introduced by Plaxco et al. [26]. We then display the ensemble of folded structures in the (CO, R_g) phase space and further calculate the potential of the mean force (PMF) using CO and R_g as reaction coordinates. Briefly, the phase space of CO and R_g is discretized into two-dimensional bins and the number of conformations within each bin, n_i , is computed. Then, the PMF value of $-k_B T \ln(n_i/n)$, where n is the total number of conformations in the ensemble, is assigned to each bin. The results are presented in a color-coded map to show the free-energy landscape.

2.9 Intramolecular contact maps

In our analysis, two beads are in contact if they are at a distance $< 7.5 \text{ \AA}$. A contact map is an array of contact-strength values defined for each pair of amino acids, i and j . The contact strength between amino acids (i, j) is equal to an average number of contacts between these two amino acids. Because there are up to four beads per amino acid, the number of contacts between i and j can be any integer between 0 and 16. The average number of contacts for each amino acid pair (i, j) is calculated by averaging the number of contacts between (i, j) over all equilibrated PGM-2X or vWF-C1 conformations.

2.10 Secondary structure analysis

We calculate the secondary structure propensities on each folded monomer conformation using the STRIDE program within the VMD software package [25]. We then average over all equilibrated PGM-2X and vWF-C conformations to obtain the average values of the α -helix, turn, and β -strand propensities per amino acid.

2.11 Surface accessible solvent area (SASA)

We calculate the coarse-grained solvent-accessible surface area (SASA) of a folded monomer by using the VMD software package [25]. In VMD, a spherical surface is created around each atom at a distance of $1.4 \text{ \AA}$, which corresponds to a single layer of water molecules away from the atom's van der Waals radius. For each residue, VMD then calculates the joint “free” surface area for all the atoms in the residue, which is the four beads and the additional backbone oxygen and hydrogen atoms that are involved in formation of the backbone hydrogen bond. The “free” surface area is defined as the area free of other atoms or bonds that would be thus exposed to the solvent. The purpose of the coarse-grained SASA defined above is to measure the free, implicit-solvent SASA available to each C $_{\beta}$ bead. Residues that are buried inside a globular folded conformation have small coarse-grained SASA values, while residues on the surface assume high coarse-grained SASA values. The coarse-grained SASA per residue thus provides information about those

residues most exposed to the solvent and is calculated first for each conformation and then averaged over all equilibrated PGM-2X or vWF-C1 conformations. In addition, we calculated the coarse-grained SASA of all hydrophobic residues first as a sum of coarse-grained SASA values over all hydrophobic residues within each conformation and then averaged over all equilibrated PGM-2X or vWF-C1 conformations.

3 Results and discussion

3.1 DMD results

We study folding of two mucin domains, PGM-2X and vWF-C1, at physiological temperature using both neutral and low pH. At neutral pH, the amino acids ASP and GLU are negatively charged and amino acids LYS and ARG are positively charged. Because only ~6% of HIS residues are charged at neutral pH, HIS is treated as a non-charged hydrophilic amino acid. At low pH, there is an excess of protons in the solvent and we allow for positively charged LYS, ARG, and HIS, whereas ASP and GLU are modeled as non-charged hydrophilic amino acids.

For each of the two protein domains, given the amino acid sequence, we first create a stretched protein chain. The desired number of distinct starting random coil conformations (32 for PGM-2X and 16 for vWF-C1, respectively) was acquired by DMD simulations performed at a high temperature of $T = 4$. The high temperature ensures that these initial conformations are random coil-like structures with zero potential energy. To quantify the conformational variability within the ensemble of initial conformations, we examined a distribution of their pairwise RMSDs for each domain (Supporting Material, Fig. S2, blue histograms). For the 242-residue-long PGM-2X, the distribution was wide, with RMSD values spread between 3 and 6 nm (Supporting Material, Fig. S2a), whereas for the 67-residue-long vWF-C1, smaller RMSD values between 1 and 2.5 nm were observed (Supporting Material, Fig. S2b) as expected for a shorter sequence. The overall high RMSD values indicate that the initial conformations were distinct and independent of each other.

We treated both implicit solvent parameters, E_{HP} and E_{CH} , as independent of pH. To test the effect of the E_{CH} parameter on the secondary structure of the folded domains, we explored a range of E_{CH} values between 0 and 1. For each value of E_{CH} , 32 (PMG-2X) and 16 (vWF-C1) different trajectories were acquired at our estimate of the physiological temperature $T = 0.13$ at each neutral and low pH, resulting in a total of $64 \times 7 = 448$ PMG-2X and $32 \times 7 = 224$ vWF-C1 folding trajectories. The trajectories at different values of E_{CH} were analyzed in terms of the average β -strand propensities, for each of the two domains at both neutral and low pH. The average β -strand propensity of vWF-C1 both at neutral and low pH did not show any clear trend with increasing E_{CH} (Fig. 2). On the other hand, the average β -strand propensity of PGM-2X at low pH considerably decreased with E_{CH} in contrast to the average β -strand propensity of PGM-2X at neutral pH, which was not sensitive to E_{CH} . Based on previous studies that estimated the electrostatic interaction strength between charged amino acids in water [24], we selected $E_{CH} = 0.8$ (or 3.68 kcal/mol) to characterize PGM-2X and vWF-C1 folding at neutral and low pH. At $E_{CH} = 0.8$, the DMD-derived folded conformations of PGM-2X were characterized by distinct average β -strand propensities of 0.26 ± 0.01 at neutral and 0.38 ± 0.01 at low pH. The vWF-C1 conformations were characterized by similar average β -strand propensities of 0.32 ± 0.02 at neutral and 0.35 ± 0.02 at low pH.

3.2 PGM-2X folding

We study folding of the 242-residue-long PGM-2X domain with a sequence [27]:

**DFNTFQNLRAKGYR~~FC~~AKPVN~~VE~~CRAESFPDTPQLQALGD~~VI~~CDKTVGLVCLNKDQ
LPPICYN~~YE~~IRILCCEMVD~~T~~CLRSTTEP~~T~~PESTRQSSSAWTPGVVSPSTQHSTATSGH
TPTASSITSSWHGPTPPSLTPCRPQCSWSK~~WF~~D~~VD~~FPPSPGPHGG~~DF~~ETYSN~~IL~~RSGEK
ICRQPEYISDLQCRAQNHPEVSIQKLGV~~VE~~CRPEVGLVCRN~~QD~~QGGKFRICLN~~YE~~V
RVLCCEPKKDCP**

Here the red letters represent charged residues and the green letters correspond to hydrophobic residues at neutral pH (according to the Kyte–Doolittle phenomenological hydropathy scale).

For comparison with the ensemble of initial conformations, the pairwise RMSD distributions within the ensemble of final conformations at neutral and low pH are shown in Fig. S2a (Supplementary Material, black and red histograms). Two distinct, partly overlapping histograms of pairwise RMSDs of final conformations for neutral and low pH, respectively, are observed. On average, the histogram of RMSD values for neutral pH is shifted towards lower values relative to the histogram of RMSD values for low pH, suggesting that PGM-2X folding is affected by pH.

We explore the conformational space by projecting each conformation onto two reaction coordinates: the contact order (CO) and radius of gyration (R_g), using distinct colors for different trajectories (Fig. 3a, b). The acquired trajectories show a significant overlap even at low pH, where a larger area of the phase space is sampled. To elucidate the free-energy landscapes at neutral and low pH, we display the PMF values for the two ensembles in the (CO, R_g) phase space (Fig. 3c, d). The PMF landscape of PGM-2X folding is rather complex. At neutral pH, a global minimum as well as several local minima are observed, suggesting the existence of multiple folding pathways. The PMF landscape undergoes significant changes upon lowering pH. At low pH, the landscape is shifted to larger R_g values and a larger number of shallower minima are found. The phase space sampled at low pH is about twofold larger in area relative to the phase space sampled at neutral pH, suggesting that lowering pH induces a transition into a more disordered peptide.

The bin with the lowest PMF value is chosen for a subsequent pairwise RMSD clustering analysis and the centroid of the largest resulting cluster is identified as a representative conformation at neutral (Fig. 4a) and low (Fig. 4b) pH. At neutral pH, the conformations are rather compact and the N- and C-termini are frequently in proximity. At low pH, a substantial increase in the β -strand structure is observed and the two termini are spatially more separated than at neutral pH. The average value of the CO at neutral pH is 0.1224 ± 0.0003 and is reduced by $\sim 18\%$ to 0.1009 ± 0.0004 at low pH. The average R_g increases from $16.66 \text{ \AA} \pm 0.02 \text{ \AA}$ at neutral to $20.71 \text{ \AA} \pm 0.04 \text{ \AA}$ at low pH.

We calculate the intramolecular contact maps of folded structures at both neutral and low pH. The results are shown in Fig. 5a, b and the contact map difference between neutral and low pH is plotted in Fig. S3 (Supporting Material). To further describe the PGM-2X folded structure we split the sequence into three regions: ASP1-THR86 (region I), PRO87-GLY158 (region II), and GLY159-PRO242 (region III), that appear on the contact maps as individual folding units with relatively few and weak interconnecting contacts (Fig. 5). Note that the choice of boundaries between these three regions is somewhat arbitrary but does not affect the conclusions.

The strongest contacts at both neutral and low pH are listed in Table 1, which shows that the dominant contacts are almost exclusively between hydrophobic amino acids. At

Fig. 3 Sampling efficiency of DMD simulations of PGM-2X folding at **a** neutral and **b** low pH and the corresponding PMF surfaces at **c** neutral and **d** low pH using the reaction coordinates CO and R_g . In **a** and **b**, each *point* corresponds to a different conformation along a trajectory and *different colors* are used for different trajectories. In **c** and **d**, we use bin sizes of 0.16% and 0.026 nm along the CO and R_g axes, respectively. The 2D plots were generated with the GNUPLOT software package

neutral pH, two out of six strongest contacts occur within region I and four out of six occur within region III. At low pH, seven out of ten strongest contacts occur within region I, two out of ten within region III and one out of ten in region II. This result indicates that the contact strengths within region I increase and those in region III decrease upon lowering pH. Regions I and III have a high density of short and long-range contacts and appear in the contact maps as sub-units, whereas region II contains mostly short-range contacts and does not form many contacts with regions I and III. Upon lowering pH, the total number and strength of contacts in region III decrease. In addition, weaker long-range contacts between the C-terminal region GLU236-PRO242 (with the sequence GLU-PRO-LYS-LYS-ASP-CYS-PRO) and regions I and II that are observed at neutral pH disappear at low pH, resulting in more extended and open structures relative to those at neutral pH. These results indicate that on lowering pH, PGM-2X partially unfolds, consistent with a shift to larger R_g values observed in Fig. 3b. In region II we observe enhanced groups of contacts that

Fig. 4 Representative conformations of a folded PGM-2X domain obtained by DMD simulations at **a** neutral and **b** low pH. First, the DMD-derived folded structures with the lowest PMF value were selected (129 at neutral and 71 at low pH). Second, the resulting structures were clustered based on their pairwise RMSD values (with cut-off 7 Å), as implemented using the GROMOS algorithm within the GROMACS software package. Third, the centroid of the largest resulting cluster was identified as a representative conformation. The N-terminal amino acid ASP1 is colored *red* and the C-terminal amino acid PRO242 is colored *blue*. The images were generated by the VMD software package [25]

are aligned perpendicular to the main diagonal, indicating an increased propensity to form β -hairpin structures contributing to an increase of the β -strand content at low pH.

Because a partial unfolding of PGM-2X is driven by pH, which affects the charged states of amino acids, we rank the strongest contacts between positively and negatively charged amino acids at neutral pH: (i) ARG169-GLU172 (with a strength of 2.7621), (ii) ARG187-GLU193 (strength 2.7153), (iii) ARG176-GLU179 (strength 2.2178) and (iv) ARG14-GLU23 (strength 2.0238). The first three contacts occur in region III and the fourth one

Fig. 5 Intramolecular contact maps of PGM-2X folded structures at **a** neutral and **b** low pH. The color scale quantifying the average number of contacts between two residues is displayed on the *right*. The SEM values are small and not visible within this color scheme

Table 1 A list of PGM-2X contacts with strengths larger than 3.5. The contacts that appear at both neutral and low pH are displayed in *bold* font. A one letter amino acid code is used here

Ranking order	Neutral pH		Low pH	
	Pair	Strength	Pair	Strength
1	L37-V41	5.3735	L37-V41	5.3775
2	A17-V20	5.3537	A17-V20	4.6767
3	A188-P192	3.8490	L199-V202	4.5365
4	L199-V203	3.6439	A10-F15	4.4749
5	L199-V202	3.6371	V74-C77	4.2209
6	L226-V230	3.5712	A36-V41	4.0693
7	—	—	V74-L78	4.0152
8	—	—	C205-V209	3.9579
9	—	—	T112-P117	3.6597
10	—	—	L8-A17	3.5217

in region I. At low pH, these contacts disappear, thus contributing to partial unfolding of region III. The drive of this unfolding can be attributed to the difference in the overall charge distribution among the three regions between the neutral and low pH. At neutral pH, region I contains 20, region II 6 and region III 24 charged amino acids with a total charge of -2 , 0 , and 0 , respectively. At a low pH, region I contains 9, region II 7, and region III 13 positively charged amino acids. Thus, an overall increase of repulsive interactions among positively charged side chains between regions I and III at low pH drives the partial unfolding of the PGM-2X domain.

The role of disulfide bonding in PGM-2X folding is elucidated by examining the probability of the disulfide bond formation as a function of separation along the sequence ([Supporting Material](#), Fig. S5a). Our results show that the disulfide bonding probability is

Fig. 6 The average **a** turn and **b** β -strand propensities per residue within PGM-2X folded structures at neutral and low pH as calculated by using the STRIDE algorithm within the VMD software package. The error bars correspond to SEM values

highest at a separation of ~ 10 – 20 residues. At neutral pH, the three strongest CYS-CYS bonds are: CYS43-CYS51 (with a strength of 1.8394), CYS51-CYS70 (strength 1.3348), and CYS225-CYS234 (strength 1.0114). At low pH, a single disulfide bond CYS51-CYS70 (strength 1.8014) contributes to folding. This result indicates that disulfide bonding contributes to PGM-2X folding mostly within region I but does not dominate the folding process.

Figure 6 shows turn and β -strand propensities per residue at both neutral and low pH. In agreement with other observations, these results indicate that at neutral pH the conformation is characterized by many turns flanked by relatively short β -strands. At low pH, a lower number of turns flanked by longer β -strands results in a reduced average turn propensity (from 0.42 ± 0.01 at neutral to 0.33 ± 0.01 at low pH) and an increased average β -strand propensity (from 0.26 ± 0.01 at neutral to 0.38 ± 0.01 at low pH). At low pH, the β -strand propensity is strongly increased in particular in region II (ARG79-THR98 and SER133-HIS157) and also in region III (VAL209-CYS234), as illustrated in Fig. 4.

We investigate the effect of pH change on the solvent-accessible surface area (SASA). The SASA per residue is calculated at neutral and low pH (Fig. 7a). To magnify the effect of pH on the SASA, the difference between SASA values at low and neutral pH is displayed in Fig. 7b, indicating the regions most affected by pH. Note that the C-terminal region GLU236-PRO242 (with the sequence GLU-PRO-LYS-LYS-ASP-CYS-PRO) is the most affected, consistent with the loss of contacts shown in Fig. 5. These results show that lowering pH results in an increased solvent exposure with the SASA increasing from

Fig. 7 **a** The average coarse-grained SASA per residue of PGM-2X folded structures at neutral and low pH. The error bars correspond to SEM values. The inset shows normalized histograms of the combined coarse-grained SASA, which is summed over all hydrophobic residues, for neutral (black) and low (red) pH. **b** The difference in coarse-grained SASA values between low and neutral pH

$39 \pm 1 \text{ \AA}^2$ to $47 \pm 2 \text{ \AA}^2$. This $\sim 21\%$ increase of the average SASA upon lowering pH may indicate a higher aggregation propensity at low pH, in particular if hydrophobic residues are more exposed. To quantify the solvent exposure of hydrophobic residues specifically, we calculate the sum of SASA values over all hydrophobic amino acids for each conformation and calculate the corresponding histograms at neutral and low pH (Fig. 7a, inset). A clear shift towards larger SASA values is observed upon lowering pH, indicating a higher solvent exposure of hydrophobic residues, which may drive mucin aggregation at low pH.

We next compare the DMD-derived conformational ensembles described above to conformations predicted by the freely available on-line protein structure prediction server I-TASSER [28, 29]. For each of the two domains, PGM-2X and vWF-C1, I-TASSER returned five different models, shown in Figs. S6 and S7 (Supporting Material), respectively. Such a comparison is limited by a small number of I-TASSER structures and by the lack of pH-specificity, which is key in our study; however, it can provide independent and valuable structural information. We assume that I-TASSER predicts a variety of structures that may be expected for a wide range of solvent conditions, including in addition to neutral also acidic pH and possibly also a membrane-mimicking environment. Of the five I-TASSER models for PGM-2X, three have a relatively compact collapsed coil tertiary structure (Fig. S6a, b, e; models 1, 2, and 5), dominated by coil, turns and short α -helices. Model 5 displays a triple β -sheet and is slightly less compact than the other two. The tertiary structure of these three models qualitatively resembles the DMD structures at a neutral pH, whereas the secondary structure of DMD-derived conformations has more β -strands and less α -helical component. As an increased amount of α -helical structure is expected to be observed in a membrane-like environment, the discrepancy in the secondary structure prediction may stem from a variety of different conditions accounted for by I-TASSER. The remaining two I-TASSER models (Fig. S6c, d; models 3 and 4) display extended conformations with either a large amount of β -sheet and no α -helical component (model 4) or small amounts of β -sheet and α -helix (model 3). These two models qualitatively resemble the PGM-2X conformations derived by the four-bead DMD approach at low pH but have a lower content of β -sheet structure than the corresponding DMD conformations.

3.3 VWF-C1 folding

The vWF-C domain in mucin is similar to the von Willebrand factor vWF-C (a protein in the blood clotting pathway) and has been shown to be involved in acid-induced polymerization. We simulated a portion of this domain, the vWF-C1 domain, which had been suggested as being involved in the pH-induced gelation of PGM [6]. The sequence of the 67-residue vWF-C1 domain [27] is:

CLGPHGEPVEPGHTVSFD CQECS CDGHTRTVSCRSQTCLPPACQEPGLVPVPEALQSGQCCPQYSC

Here, residues, which are charged at neutral pH, are marked by red letters and hydrophobic residues are marked by green letters.

In the case of vWF-C1 the initial and final pairwise RMSD distributions are shown in Fig. S2b (Supporting Material, black and red histograms). Two overlapping histograms of pairwise RMSDs of final conformations for neutral and low pH, respectively, are

Fig. 8 Sampling efficiency of DMD simulations of vWF-C1 folding at **a** neutral and **b** low pH and the corresponding PMF surfaces at **c** neutral and **d** low pH using the reaction coordinates CO and R_g . In **a** and **b**, each *point* corresponds to a different conformation along a trajectory and *different colors* are used for different trajectories. In **c** and **d**, we use bin sizes of 0.20% and 0.007 nm along the CO and R_g axes, respectively. The 2D plots were generated with the GNUPLOT software package

observed, suggesting no significant pH-induced change in the overall size of the folded structure.

The conformational space of vWF-C1 is displayed in Fig. 8a, b indicating a significant overlap between the trajectories at both neutral and low pH. The PMF landscape (Fig. 8c, d) is characterized by a single minimum at both neutral at low pH. In contrast to the PGM-2X folding landscape, the vWF-C1 folding landscape is characterized by a single broad free-energy minimum, which at low pH becomes more compact than at neutral pH.

Representative conformation for vWF-C1 at neutral (Fig. 9a) and low (Fig. 9b) pH were identified using the method described for PGM-2X. The average value of the CO at neutral pH is 0.1748 ± 0.0005 and increases to 0.1856 ± 0.0005 at low pH. The average radius of gyration, R_g , increases from 9.84 ± 0.01 Å at neutral to 9.95 ± 0.01 Å at low pH. Both the changes in CO and R_g are relatively minor compared to the corresponding pH-induced changes in PGM-2X. Interestingly, the average CO for the ensemble of folded structures is

Fig. 9 Representative conformations of a folded vWF-C1 domain obtained by DMD simulations at **a** neutral and **b** low pH. First, the DMD-derived folded structures with the lowest PMF value were selected (19 at neutral and 19 at low pH). Second, the resulting structures were clustered based on their pairwise RMSD values (with cut-off 7 Å), as implemented using the GROMOS algorithm within the GROMACS software package [36, 37]. Third, the centroid of the largest resulting cluster was identified as a representative conformation. The N-terminal amino acid CYS1 is colored *red* and the C-terminal amino acid CYS67 is colored *blue*. The images were generated by the VMD software package [25]

higher for vWF-C1 than for PGM-2X and *slightly increases* upon lowering pH, indicating that low pH does not cause vWF-C1 unfolding but rather stabilizes the folded structure.

Intramolecular contact maps for vWF-C1 at both neutral and low pH are shown in Fig. 10a, b. The contact map difference between neutral and low pH is plotted in Fig. S4 (Supporting Material). The strongest contacts at both neutral and low pH are listed in

Fig. 10 Intramolecular contact maps of vWF-C1 folded structures at **a** neutral and **b** low pH. The color scale quantifying the average number of contacts between two residues is displayed on the *right*. The SEM values are small and not visible within this color scheme

Table 2 A list of vWF-C1 contacts with strengths larger than 3.5. The contacts that appear at both neutral and low pH are displayed in *bold* font. A one letter amino acid code is used here

Ranking order	Neutral pH		Low pH	
	Pair	Strength	Pair	Strength
1	L56-C61	5.3082	L56-C61	6.5990
2	C1-V9	4.9548	C24-R29	5.7271
3	C24-R29	4.7290	C22-V31	4.8490
4	V9-V15	4.3199	L52-A55	4.2710
5	L2-V9	4.3063	L40-A43	4.1300
6	L50-L52	3.9443	L2-V9	3.8973
7	L40-A43	3.8601	C1-V9	3.6244
8	C10-F17	3.6937	—	—

Table 2 and indicate, as in the case of PGM-2X, that the dominant contacts are mostly between hydrophobic amino acids. The strongest contact at both neutral and low pH is LEU56-CYS61, which is central to the folded structure at the C-terminal region. The contact between a hydrophobic and charged residue, CYS24-ARG29, which is strong at both neutral (with a strength of 4.7290) and low (with a strength of 5.7271) pH, is a central contact around which a β -hairpin structure centered at GLY26 is formed. At neutral pH, this contact is enhanced by a contact between two oppositely charged residues (ASP25-ARG29 with a strength of 2.7314), whereas at low pH it is enhanced by a contact between two hydrophobic residues (CYS22-VAL31 with a strength of 4.8490). This β -hairpin structure spans residues GLU21-SER32 (as well as some surrounding amino acids) and represents one of the main features of the vWF-C1 folded structure. The region CYS1-LEU2 strongly interacts with VAL9 and also forms long-range contacts along the entire sequence. It is important to note that at low pH, the number of long-range contacts between the CYS1-LEU2 region and the rest of the peptide *increases*, thus contributing to an increase of the average CO, which is dominated by the presence of long-range contacts. Even though a few changes in the average folded structure are observed upon lowering pH (see also Fig. S4 in [Supplementary Material](#)), the main features remain preserved. Thus, under the same conditions, the effect of pH on the two mucin domains is quite different.

At neutral pH, the total charge of the vWF-C1 domain is -5 and, at low pH, the total charge is $+5$. We thus examine whether effective attractive interactions between oppositely charged residues contribute to folding at neutral pH. Our analysis shows that the only contact between oppositely charged residues that occurs at neutral pH, ASP25-ARG29, has a strength of 2.7314, which is not among the strongest contacts. Even though this contact contributes to the β -hairpin structure centered at GLY26 at neutral pH, its absence at low pH does not strongly destabilize this β -hairpin structure. Rather, the effective repulsive interactions among the positively charged residues at low pH result in a slight re-distribution of contacts observed in Fig. 10a, b.

The vWF-C1 sequence contains 10 CYS amino acids. The probability of disulfide bond formation as a function of separation along the sequence for vWF-C1 is shown in Fig. S5b ([Supporting Material](#)) and indicates that the disulfide bonding probability is highest at a separation of $\sim 4-6$, followed by that at separation of $\sim 10-20$ residues. Note that the overall probabilities for disulfide bond formation are somewhat larger for vWF-C1 than for PGM-2X, which is consistent with a larger density of CYS residues in the sequence. At neutral pH, the strongest CYS-CYS bonds are CYS1-CYS19 (with a strength of 3.4084) and CYS19-CYS33 (with a strength of 2.2036), whereas at low pH, a single disulfide bond CYS19-CYS22 (with a strength of 2.6170) contributes to folding. None of the disulfide bonds can be found among the overall strongest contacts (Table 2). This result indicates that although

Fig. 11 The average **a** turn and **b** β -strand propensities per residue within vWF-C1 folded structures at neutral and low pH as calculated by using the STRIDE algorithm within the VMD software package. The error bars correspond to SEM values

disulfide bonding contributes to vWF-C1 folding, it does not dominate the folding process itself.

The turn and β -strand propensities per residue at both neutral and low pH are shown in Fig. 11 and indicate that for both pH values the folded structure is characterized by several turns flanked by β -strands. This is in agreement with the distribution of contacts in Fig. 10. The number of turns and their positions are not affected by pH, nor is the average turn propensity, which is 0.38 ± 0.03 at neutral and 0.37 ± 0.03 at low pH (Fig. 11a). A similar conclusion is made for the β -strand propensity. Although in a few regions the β -strand propensity decreases upon lowering pH, it increases in others, such that at neutral pH the average β -strand propensity is 0.32 ± 0.02 and at low pH 0.35 ± 0.02 , a difference that is not statistically significant.

Finally, the SASA per residue is calculated at neutral and low pH (Fig. 12a). The difference between SASA values at low and neutral pH is displayed in Fig. 12b, showing the residues that are most affected by pH. Note that at low pH a few regions, including the C-terminal region TYR65-CYS67, are more exposed to the solvent than at neutral pH. The largest changes (peaks in Fig. 12b) correspond to charged residues. On average, however, the total SASA increased from $50 \pm 3 \text{ \AA}^2$ to $52 \pm 3 \text{ \AA}^2$, which is not statistically significant. To quantify the solvent exposure of only hydrophobic residues, we calculate the

Fig. 12 **a** The average coarse-grained SASA per residue of vWF-C1 folded structures at neutral and low pH. The *error bars* correspond to SEM values. The *inset* shows normalized histograms of the combined coarse-grained SASA, which is summed over all hydrophobic residues, for neutral (*black*) and low (*red*) pH. **b** The difference in coarse-grained SASA values between low and neutral pH

sum of SASA values over all hydrophobic amino acids for each conformation and calculate the corresponding histograms at neutral and low pH (Fig. 12a, inset). A slight (but not statistically significant) shift towards smaller SASA values is observed upon lowering pH, suggesting that the aggregation propensity of the vWF-C1 domain may not be strongly affected by pH. A plausible mechanism of quaternary interactions among the vWF-C1 domains at low pH may include disulfide bonding involving CYS67, which undergoes the largest pH-induced change in the SASA and is a residue with highest solvent exposure at low pH.

For comparison, we included the I-TASSER predictions resulting in five models of vWF-C1 folded structures, again regardless of the solvent conditions. Snapshots of the five predicted models are shown in Fig. S7a–e (Supporting Material). The vWF-C1 folded structure models predicted by I-TASSER look qualitatively similar to the representative structures obtained by the four-bead DMD approach in terms of the secondary structure, dominated by coil, turns and β -strands but have a less compact tertiary structure than the DMD conformations. Interestingly, the five I-TASSER conformations are similar in size

and content, in agreement with our four-bead DMD observation that pH does not affect the compactness or the secondary structure of the vWF-C1 conformations.

4 Conclusions

The DMD approach combined with a four-bead peptide model in an implicit solvent was designed to study protein folding and assembly for large systems and on a time scale not accessible to all-atom MD simulations. Although the approach was mostly applied to studies of Alzheimer's amyloid β -protein folding and oligomer formation [15, 16, 18, 19, 23, 30] and recently validated by all-atom MD in an explicit solvent [31], the parameters of the model that are based on the known structure of many proteins from Protein Data Bank (<http://www.rcsb.org/pdb/>) are not protein-specific.

In the present work, DMD simulations of PGM-2X and vWF-C1 folding using a four-bead protein model with hydrogen bonding and residue-specific effective hydrophobic and electrostatic interactions reveal novel structural information in relation to the aggregation affinity of the two protein domains at neutral and low pH. The folded structure of both domains consists of β -strands separated by several turns and loops at both neutral and low pH. PGM-2X exhibits a conformational transition to a more open, extended, and disordered structure upon lowering pH, accompanied by a decrease of the contact order and increase of the average β -strand content. In contrast, the globular conformation of vWF-C1 is not strongly influenced by pH, although some structural rearrangement takes place and the C-terminal residue CYS67 becomes significantly more exposed to the solvent upon lowering pH. The contact order in vWF-C1 increases upon lowering pH, in contrast to PGM-2X. The DMD results were derived by using the same set of implicit solvent parameters for both the PGM-2X and vWF-C1. The simulations predict an increase in β -strand for PGM-2X upon lowering pH, while that of the vWF-C1 domain is unaffected by pH. Examination of free-energy landscapes at neutral pH demonstrates that PGM-2X has a complex folding landscape with multiple minima in contrast to vWF-C1 with a single free-energy minimum. While stabilizing the folded structure of the vWF-C1 domain, low pH induces a significant change in the free-energy landscape of the PGM-2X domain with an increased number of shallower minima, resembling folding landscapes of intrinsically disordered proteins associated with aberrant aggregation and human pathology.

DMD simulations with the four-bead peptide model show that folding of both PGM-2X and vWF-C1 is dominated by contacts among hydrophobic residues. This observation is consistent with recently reported Monte Carlo simulations of a sol-gel transition in a coarse-grained model (20-bead) of mucus, which demonstrated a critical role of effective hydrophobic interactions whereas disulfide bonds were shown to affect the frictional properties of mucus [32]. Our DMD approach using a model with four beads per amino acid is of intermediate resolution with amino acid-specific details and thus provides a more detailed structural information complementary to the above computational study. Our simulations show that lowering pH results in a partial unfolding of PGM-2X caused by a loss of long-range contacts between the C-terminal region GLU236-PRO242 and the rest of the domain. This seven-residue C-terminal region contains 4 charged residues as well as CYS residues, which at low pH are significantly more exposed to the solvent than at neutral pH. PGM-2X folded structures reveal three peptide regions: N-terminal, central, and C-terminal, which appear to fold almost independently of each other. Inspection of contacts between oppositely charged residues does not reveal any specific contact that dominates

the process. Instead, the unfolding seems to be a consequence of “mean-field” repulsive interactions between strongly positively charged N-terminal and C-terminal regions at low pH. This partial unfolding is associated with an increased solvent exposure of the entire PGM-2X domain and in particular its hydrophobic residues, which may contribute to PGM-2X aggregation propensity at low pH. These results support the model for gelation of mucin developed by Cao et al. [6] based on their observation of increased binding of hydrophobic dyes at pH below 4. Cao et al. suggested that gelation involves the interplay of electrostatic and hydrophobic interactions of non-glycosylated regions of the mucin polymer. At neutral pH, hydrophobic domains are sequestered in folds stabilized by salt bridges between negatively charged carboxylates of ASP or GLU (pK around 4) and positively charged amino groups of LYS, ARG, or HIS. At acidic pH, the protonation of the COOH group breaks salt bridges, thereby unfolding and exposing hydrophobic regions of protein that can associate acting as the crosslinks of a gel. The results of the present DMD simulations further suggest that PGM-2X, in particular its C-terminal region, is likely to be involved in the gelation of gastric mucin. The folding of the vWF-C1 domain may also be of relevance in understanding the structure of the von Willebrand factor, the protein involved in the aggregation of platelets in the blood clotting process [33, 34].

Acknowledgements The authors thank Dr. Yuriy V. Sereda for his contribution to the implementation of PRO amino acid into the four-bead protein model. B.U. and B.B. acknowledge the support by the NIH grant AG027818 and thank NSF for the access to the Extreme Science and Engineering Discovery Environment (XSEDE) supercomputing facilities through the grant PHY100030. B.S.T. thanks Dr. Nezam Afdhal, M.D., Beth Israel Deaconess Medical Center for financial support.

References

1. Bansil, R., Turner, B.S.: Mucin structure, aggregation, physiological functions and biomedical applications. *Curr. Opin. Colloid Interface Sci.* **11**, 164–170 (2006)
2. Strous, G.J., Dekker, J.: Mucin-type glycoproteins. *Crit. Rev. Biochem. Mol. Biol.* **27**, 57–92 (1992)
3. Perez-Vilar, J., Hill, R.L.: The structure and assembly of secreted mucins. *J. Biol. Chem.* **274**, 31751–31754 (1999)
4. Turner, B.S.: Expression of cysteine-rich pig gastric mucin (MUC5AC) domains in *Pichia pastoris* and their pH-dependent properties. Dissertation, Boston University (2012)
5. Bhaskar, K.R., Gong, D., Bansil, R., Pajevic, S., Hamilton, J.A., Turner, B.S., Lamont, J.T.: Profound increase in viscosity and aggregation of pig gastric mucin at low pH. *Am. J. Physiol.* **261**, G827–G833 (1991)
6. Cao, X.X., Bansil, R., Bhaskar, K.R., Turner, B.S., LaMont, J.T., Niu, N., Afdhal, N.H.: pH-dependent conformational change of gastric mucin leads to sol-gel transition. *Biophys. J.* **76**, 1250–1258 (1999)
7. Celli, J.P., Turner, B.S., Afdhal, N.H., Ewoldt, R.H., McKinley, G.H., Bansil, R., Erramilli, S.: Rheology of gastric mucin exhibits a pH-dependent sol-gel transition. *Biomacromolecules* **8**, 1580–1586 (2007)
8. Hollander, F.: The two-component mucous barrier; its activity in protecting the gastroduodenal mucosa against peptic ulceration. *AMA Arch. Intern. Med.* **93**, 107–120 (1954)
9. Hong, Z., Chasan, B., Bansil, R., Turner, B.S., Bhaskar, K.R., Afdhal, N.H.: Atomic force microscopy reveals aggregation of gastric mucin at low pH. *Biomacromolecules* **6**, 3458–3466 (2005)
10. Cao, X.: Aggregation and gelation of mucin and its interactions with lipid vesicles. Dissertation, Boston University (1997)
11. Rapaport, D.C.: *The Art of Molecular Dynamics Simulation*. Cambridge University Press, Cambridge (1995)
12. Teplow, D.B., Lazo, N.D., Bitan, G., Bernstein, S., Wyttenbach, T., Bowers, M.T., Baumketner, A., Shea, J.E., Urbanc, B., Cruz, L., Borreguero, J., Stanley, H.E.: Elucidating amyloid β -protein folding and assembly: a multidisciplinary approach. *Acc. Chem. Res.* **39**, 635–645 (2006)
13. Smith, A.V., Hall, C.K.: α -helix formation: discontinuous molecular dynamics on an intermediate-resolution protein model. *Protein Struct. Funct. Genet.* **44**, 344–360 (2001)

14. Ding, F., Borreguero, J.M., Buldyrev, S.V., Stanley, H.E., Dokholyan, N.V.: Mechanism for the α -helix to β -hairpin transition. *Proteins* **53**, 220–228 (2003)
15. Urbanc, B., Borreguero, J.M., Cruz, L., Stanley, H.E.: *Ab initio* discrete molecular dynamics approach to protein folding and aggregation. *Meth. Enzymol.* **412**, 314–338 (2006)
16. Urbanc, B., Cruz, L., Yun, S., Buldyrev, S.V., Bitan, G., Teplow, D.B., Stanley, H.E.: In silico study of amyloid β -protein folding and oligomerization. *Proc. Natl. Acad. Sci. USA* **101**, 17345–17350 (2004)
17. Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982)
18. Lam, A.R., Teplow, D.B., Stanley, H.E., Urbanc, B.: Effects of the Arctic (E22→G) mutation on amyloid β -protein folding: discrete molecular dynamics study. *J. Am. Chem. Soc.* **130**, 17413–17422 (2008)
19. Urbanc, B., Betnel, M., Cruz, L., Bitan, G., Teplow, D.B.: Elucidation of amyloid β -protein oligomerization mechanisms: discrete molecular dynamics study. *J. Am. Chem. Soc.* **132**, 4266–4280 (2010)
20. Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., Dinola, A., Haak, J.R.: Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984)
21. Nosé, S.: A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **81**, 511 (1984)
22. Dani, V.S., Ramakrishnan, C., Varadarajan, R.: MODIP revisited: re-evaluation and refinement of an automated procedure for modeling of disulfide bonds in proteins. *Protein Eng.* **16**, 187–193 (2003)
23. Urbanc, B., Betnel, M., Cruz, L., Li, H., Fradinger, E.A., Monien, B.H., Bitan, G.: Structural basis for A β 1–42 toxicity inhibition by A β C-terminal fragments: discrete molecular dynamics study. *J. Mol. Biol.* **410**, 316–328 (2011)
24. Masunov, A., Lazaridis, T.: Potentials of mean force between ionizable amino acid side chains in water. *J. Am. Chem. Soc.* **125**, 1722–1730 (2003)
25. Humphrey, W., Dalke, A., Schulten, K.: VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996)
26. Plaxco, K.W., Simons, K.T., Baker, D.: Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994 (1998)
27. Turner, B.S., Bhaskar, K.R., Hadzopoulou-Cladaras, M., LaMont, J.T.: Cysteine-rich regions of pig gastric mucin contain von Willebrand factor and cystine knot domains at the carboxyl terminal. *Biochim. Biophys. Acta* **1447**, 77–92 (1999)
28. Roy, A., Kucukural, A., Zhang, Y.: I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**, 725–738 (2010)
29. Zhang, Y.: Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* **69**(Suppl 8), 108–117 (2007)
30. Yun, S., Urbanc, B., Cruz, L., Bitan, G., Teplow, D.B., Stanley, H.E.: Role of electrostatic interactions in amyloid β -protein (A β) oligomer formation: a discrete molecular dynamics study. *Biophys. J.* **92**, 4064–4077 (2007)
31. Barz, B., Urbanc, B.: Dimer formation enhances structural differences between amyloid β -protein (1–40) and (1–42): an explicit-solvent molecular dynamics study. *PLoS ONE* **7**, e34345 (2012)
32. Gniewek, P., Kolinski, A.: Coarse-grained Monte Carlo simulations of mucus: structure, dynamics, and thermodynamics. *Biophys. J.* **99**, 3507–3516 (2010)
33. Ruggeri, Z.M.: von Willebrand factor. *J. Clin. Invest.* **100**, S41–S46 (1997)
34. de Wit, T.R., van Mourik, J.A.: Biosynthesis, processing and secretion of von Willebrand factor: biological implications. *Best. Pract. Res. Cl. Ha.* **14**, 241–255 (2001)
35. Dekker, J., Rossen, J.W.A., Buller, H.A., Einerhand, A.W.C.: The MUC family: an obituary. *Trends Biochem. Sci.* **27**, 126–131 (2002)
36. Berendsen, H.J.C., van der Spoel, D., van Drunen, R.: GROMACS: a message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91**, 43–56 (1995)
37. Lindahl, E., Hess, B., van der Spoel, D.: GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* **7**, 306–317 (2001)